

# UC San Diego

## UC San Diego Previously Published Works

### Title

Microbial Similarity between Students in a Common Dormitory Environment Reveals the Forensic Potential of Individual Microbial Signatures.

### Permalink

<https://escholarship.org/uc/item/5mp1q12v>

### Journal

mBio, 10(4)

### ISSN

2150-7511

### Authors

Richardson, Miles  
Gottel, Neil  
Gilbert, Jack A  
et al.

### Publication Date

2019-07-01

### DOI

10.1128/mbio.01054-19

Peer reviewed



# Microbial Similarity between Students in a Common Dormitory Environment Reveals the Forensic Potential of Individual Microbial Signatures

 Miles Richardson,<sup>a,b</sup> Neil Gottel,<sup>c</sup> Jack A. Gilbert,<sup>c</sup> Simon Lax<sup>d</sup>

<sup>a</sup>Department of Systems Biology, Columbia University, New York, New York, USA

<sup>b</sup>Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, New York, USA

<sup>c</sup>Department of Pediatrics, University of California San Diego, La Jolla, California, USA

<sup>d</sup>Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

**ABSTRACT** The microbiota of the built environment is an amalgamation of both human and environmental sources. While human sources have been examined within single-family households or in public environments, it is unclear what effect a large number of cohabitating people have on the microbial communities of their shared environment. We sampled the public and private spaces of a college dormitory, disentangling individual microbial signatures and their impact on the microbiota of common spaces. We compared multiple methods for marker gene sequence clustering and found that minimum entropy decomposition (MED) was best able to distinguish between the microbial signatures of different individuals and was able to uncover more discriminative taxa across all taxonomic groups. Further, weighted UniFrac- and random forest-based graph analyses uncovered two distinct spheres of hand- or shoe-associated samples. Using graph-based clustering, we identified spheres of interaction and found that connection between these clusters was enriched for hands, implicating them as a primary means of transmission. In contrast, shoe-associated samples were found to be freely interacting, with individual shoes more connected to each other than to the floors they interact with. Individual interactions were highly dynamic, with groups of samples originating from individuals clustering freely with samples from other individuals, while all floor and shoe samples consistently clustered together.

**IMPORTANCE** Humans leave behind a microbial trail, regardless of intention. This may allow for the identification of individuals based on the “microbial signatures” they shed in built environments. In a shared living environment, these trails intersect, and through interaction with common surfaces may become homogenized, potentially confounding our ability to link individuals to their associated microbiota. We sought to understand the factors that influence the mixing of individual signatures and how best to process sequencing data to best tease apart these signatures.

**KEYWORDS** built environments, microbial ecology, microbial transmission

Numerous recent studies have uncovered the extent to which humans influence the microbial ecology of the spaces they occupy through microbial exchange between skin and the built environment. Most of these studies have focused on home-associated microbial communities (1–4), with home size, number of occupants, and building materials differentiated between sampling locations. Each of those confounding factors may have a significant impact on microbial community structure, and they are difficult to disentangle. Other studies have focused instead on the microbial ecology of public spaces, such as classrooms and hospital entrance halls (5–11). Although they have been

**Citation** Richardson M, Gottel N, Gilbert JA, Lax S. 2019. Microbial similarity between students in a common dormitory environment reveals the forensic potential of individual microbial signatures. *mBio* 10:e01054-19. <https://doi.org/10.1128/mBio.01054-19>.

**Editor** Mark J. Bailey, CEH-Oxford

**Copyright** © 2019 Richardson et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Miles Richardson, [miles.richardson@columbia.edu](mailto:miles.richardson@columbia.edu), or Simon Lax, [simonlax@mit.edu](mailto:simonlax@mit.edu).

**Received** 25 April 2019

**Accepted** 1 July 2019

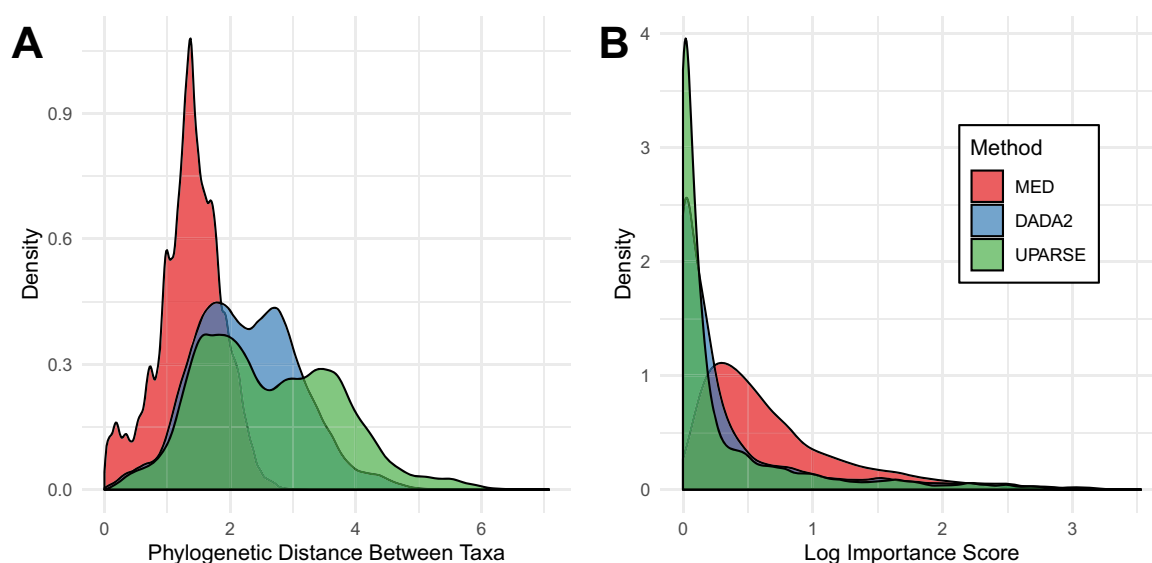
**Published** 30 July 2019

able to demonstrate that most of the taxa colonizing those spaces are skin-associated taxa, they are unable to link individual human microbial signatures to their data.

Individuals create their own “microbial cloud” (12) by constantly shedding their own microbiota (13, 14). Individuals shed around 30 million bacterial cells per hour (13), and thus leave behind a “microbial fingerprint” which has been shown to be stable over time (15, 16), although body sites vary in their stability (17). Microbial flow in the built environment is a keen topic of interest, as human skin is the dominant contributor to the microbiome of built environments (18). Cohabitation of multiple individuals has been shown to influence the microbiota of common spaces and of the constituents themselves (3, 8, 11), and common areas may serve as mechanisms of microbial exchange between individuals (19, 20). Dormitory buildings, which have a standardized architectural design, common building materials and furnishings in the rooms, and even a common ventilation system, represent an intriguing model system in which to characterize the direct effects of an individual’s skin microbiota on their surroundings and to further elucidate the forensic potential of skin microbial signatures. In one sense, dorm rooms represent a number of replicates that can be used to uncover general patterns of human microbial exchange with the built environment. In a different sense, they are a “metacommunity” in which it is possible to record a network of interaction by logging visits between rooms and the use of common spaces. The divide between private rooms and common spaces such as hallways, lounges, and restrooms further enables us to tease apart individual microbial signatures in shared spaces.

Identifying microbial signatures relies on recovering individual-specific taxa, either through the use of universal markers such as the 16S gene (21), clade-specific markers (22, 23), or metagenomic information (24). It is unclear how methodological differences in sequence clustering impact the ability to link individuals to their surroundings through microbial similarity. To determine how to optimize the inference of individual microbial signatures, we employed three sequence processing methods to determine which was most discriminative in characterizing individuals. It has been observed that in many built environment studies, a large fraction of reads are attributed to a small number of operational taxonomic units (OTUs) (3, 11, 25). These OTUs come from a small selection of skin-associated taxonomic groups, including corynebacteria, staphylococci, pseudomonads, and streptococci (26, 27). As much of the differentiation between individuals occurs within a small number of taxonomic groups, it is unclear how to optimize sequence clustering for forensic inference, as OTU clustering may lump together similar sequences by design. OTU clustering is commonly used as a way to control for error introduced during sequence processing and sequencing, which can produce artifacts that obscure the true composition of a sample (28, 29). OTU clustering commonly occurs at the 97% similarity level, as this roughly corresponds to species identity (30). Among the most commonly used pipelines for OTU clustering is UPARSE (31), a method that constructs OTUs by prioritizing highly abundant unique sequences during clustering, as these highly abundant sequences are less likely to be sequencing errors.

At the same time, OTU clustering has limits, as the 97% threshold erases significant differences within closely related taxa and can overestimate similarity between taxa (32, 33). To overcome these limits, a number of methods have been introduced to determine exact sequences without clustering (34–37). DADA2 (36) is a reference-free sequence-based algorithm that separates sequence errors from biological variation based on a Poisson error model, which partitions reads into unique sequences by grouping together sequences with a high likelihood of sequencing error. In contrast, minimum entropy decomposition (MED) (35) is an unsupervised version of oligotyping (34), a method that iteratively partitions sequences based on Shannon entropy. A sequence alignment is generated, and nucleotide sites with high nucleotide variation are used to partition groups of sequences, which then proceeds iteratively within each partition until no sites contain sufficient variation to merit further decomposition. Oligotyping has been used to explore variation in host-associated bacteria (38, 39) and uncover bacterial ecotypes (40, 41). The increased ability to discriminate between



**FIG 1** (A) Distribution of phylogenetic distance, based on the pairwise phylogenetic branch length between all taxa by each sequence processing method. MED recovers more highly related taxa than DADA2 or UPARSE. (B) Distribution of importance scores over all taxa, grouped by sequence processing method. The y axis is log transformed to aid visualization.

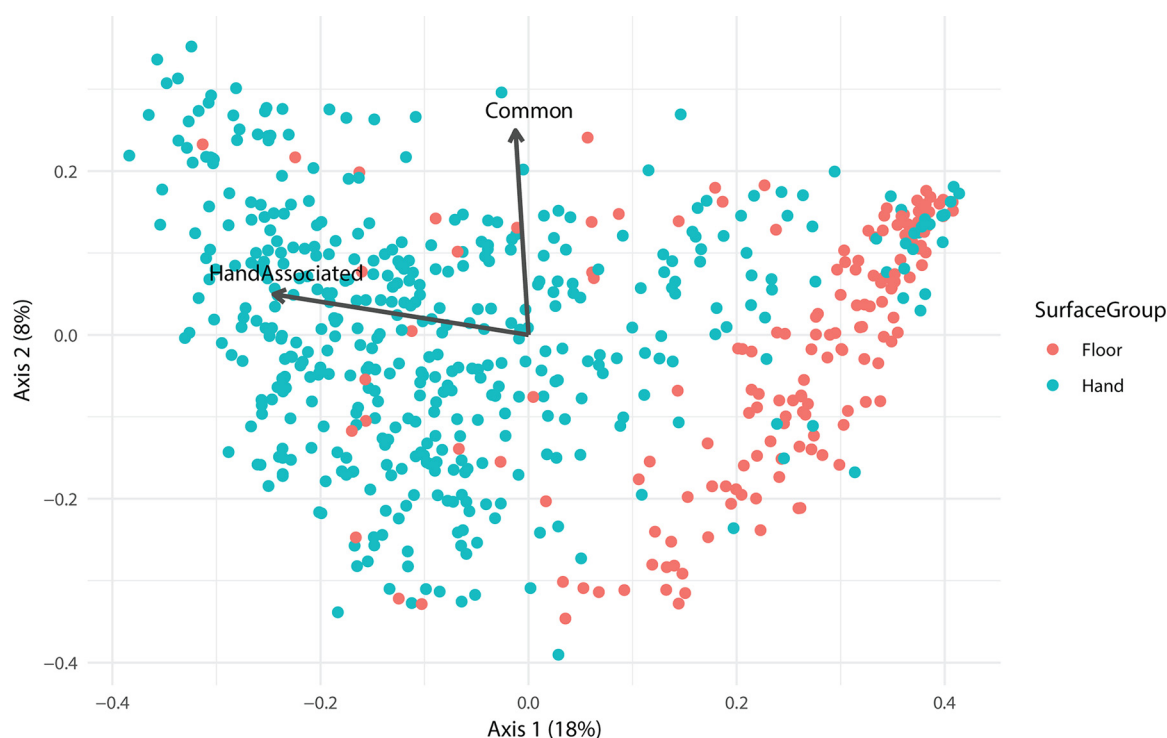
closely related taxa, such as in *Blautia* found in sewage systems (42), allows for the identification of their hosts.

To explore the divide between public and private, we sampled 37 participants and their rooms from floors five through eight of the University of Chicago's eight floor South Campus residence hall, with four time points over 3 months. Participants were drawn from one "house" in the dormitory, which serves a subset of the dormitory floor plan with shared common space and bathrooms. From participants, we swabbed both the skin of their dominant hand, and their personal effects, such as bed sheets and shoes. Additionally, common surfaces on each floor, including tables and bathrooms, were also sampled. Dorm rooms and common spaces had openable windows, along with forced air heating and cooling. Together, this collection of surfaces encompasses the divide between private and public space in the dormitory. Further, to determine the most effective 16S rRNA sequence processing method to enable individual identification, we employed UPARSE at 97% identity, DADA2, and MED on these collected data.

## RESULTS

**Clustering methodology impacts the success of forensic inference.** Each of the sequence processing methods produced a different picture of the microbial diversity of the dormitory. UPARSE recovered the largest number of distinct sequences (6,011) along with the greatest number of phyla (25 phyla). DADA2 recovered nearly the same phylum level diversity as UPARSE (23 versus 25) but fewer sequences (4,307). MED recovered fewer sequences (3,353) and fewer phyla (9 phyla) but recovered more members within each phylum (see Table S1 in the supplemental material). MED also had a significantly smaller phylogenetic distance between taxa (Wilcoxon rank sum test,  $P < 2.2 \times 10^{-16}$ ) than both DADA2 and UPARSE (Fig. 1A), indicating that MED recovered much more closely related sequences.

Since we were most interested in classifying individuals, we compared each method using a random forest model trained on surfaces that closely associate with the hands of only one individual in order to test their forensic inference. There is a major divide between floor- and hand-associated samples (Fig. 2). Floor-associated samples, including shoes and floors, inhabit a different space compared to hand-associated samples, and this division significantly structures these communities (analysis of similarities [ANOSIM] on Bray-Curtis distance,  $R = 0.2821$ ,  $P = 0.001$ ). Thus, to predict which indi-



**FIG 2** A principal component analysis (PCoA) plot based upon the Bray-Curtis distance. Statistically significant environmental vectors ( $P < 0.01$ ) by *envfit* are plotted over the data. Common surface ( $R^2 = 0.0654$ ,  $P < 10^{-6}$ ) and hand-associated ( $R^2 = 0.38$ ,  $P < 10^{-6}$ ) vectors are shown.

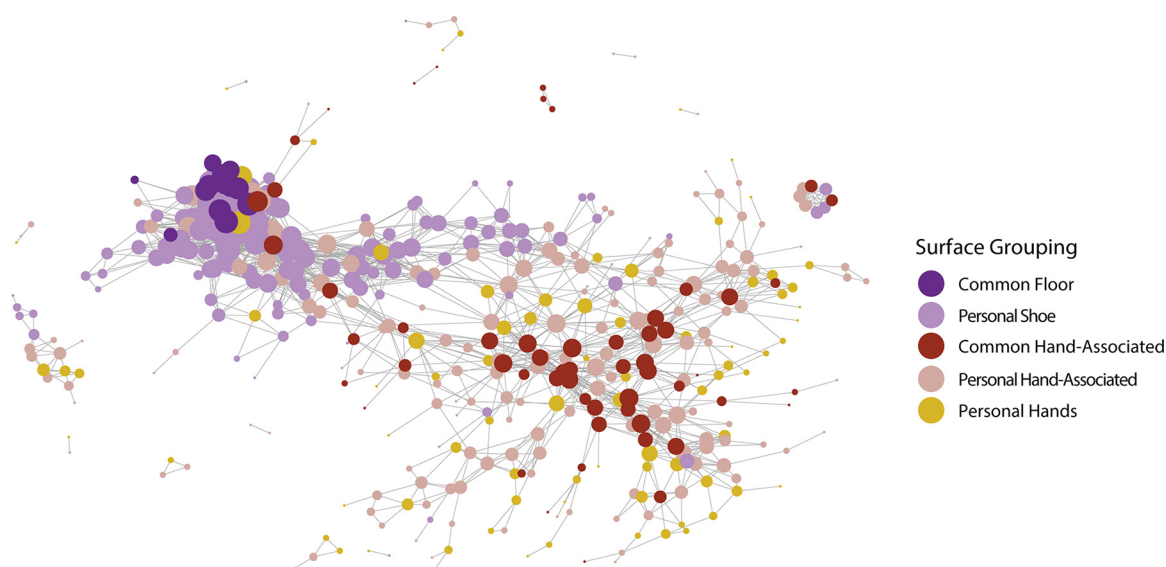
vidual's hands a surface had interacted with, bed sheets, desks, and door handles of the participant rooms are the most useful.

These models were implemented using a random forest model (43), which allows for the interrogation of similarity between samples. The model was then tested on hand samples from the same individuals, with the resulting accuracy summarized in Table 1. The standardized method of interpreting the success of classifiers is the error ratio, which quantifies how well the random forest model does at predicting the correct individual relative to the success expected by chance (44). An error ratio above two is commonly used as a significance threshold, and a higher ratio indicates better performance. All methods performed significantly better than random, but MED clearly outperformed UPARSE and DADA2 in our data set. Figure S1 in the supplemental material presents the confusion matrix generated by MED. Samples that fall on the diagonal are correctly classified by the random forest model. Most samples (79.57%) fall on the diagonal of the plot. However, for certain individuals, their hand samples are misclassified in every instance.

Interestingly, the largest source of classification error was the presence of roommates in the room. In fact, the classification error of an individual was linearly related to the number of roommates that individual had ( $R^2 = 0.3143$ ,  $P < 0.0001$ ), with classification error increasing by 18 percentage points for each additional roommate. The relationship is shown in Fig. S2. The random forest model attempts to use differences in taxon abundance between individuals to classify individuals. If two individuals interact and exchange bacteria, differences in abundance decrease, which in

**TABLE 1** Random forest model accuracy and error ratios

Method	Accuracy (CV-10) (%)	Error ratio
UPARSE	60.96	2.49
DADA2	71.06	3.36
MED	79.57	4.76



**FIG 3** A weighted UniFrac graph of all samples, thresholded to be below 0.12 weighted UniFrac distance between individuals. They are sized based on their degree centrality, a measure of the number of connections they have to other samples. Samples are colored by sample type, with desks, bed sheets, and door handles grouped together as personal hand-associated samples. Common hand-associated surfaces act as a scaffold, connecting between themselves, along with connecting many distinct individuals.

turn increases model error. Roommates had a significantly smaller weighted UniFrac distance between them than individuals residing in different rooms (Wilcoxon rank sum test,  $W = 409660000$ ,  $P < 2.2 \times 10^{-16}$ ).

**Classification of individuals is driven by specific taxa.** The random forest model is able to rank individual sequences or OTUs by their importance to successful classification. During the random forest generation process, only two thirds of variables are used to generate each forest. The accuracy of forests containing a given bacterial sequence can be compared to those without the sequence, and this is used to calculate the importance score. MED recovers significantly higher importance scores than DADA2 or UPARSE and has a distinct distribution as seen in Fig. 1B. Furthermore, MED has a significantly higher average importance score (Wilcoxon rank sum test, false-discovery rate [FDR]  $P < 0.05$ ) (Fig. S3) across all phyla that overlap between all three methods except for *Cyanobacteria*, *Fusobacteria*, and *Deinococcus-Thermus*.

It has been noted that there are taxa indicative of different sexes (45). To see whether there were enriched taxa between men and women from room samples, we looked for differentially enriched taxa using DESeq2. The most significantly enriched taxon was *Lactobacillus iners*, an inhabitant of the female reproductive tract. Certain corynebacteria were also noted to be more abundant in men, as seen in Fig. S4. Using these enriched taxa, we used the random forest model to predict whether a subject is a man or woman, with an error ratio of about 2.5, and accuracy of around 80% on the test set.

**Metacomunity structure.** In addition to classifying individuals, we sought to recapitulate the geographical structure of the dorm using graphical models. To do this, we constructed a threshold graph of the weighted UniFrac distance between samples, with edges preserved if they were less than a threshold of 0.12. As seen in Fig. 3, the dorm has two large subgraphs, along with a number of orphaned graphs. These two groups consist of floor-associated (shoes and floors) and hand-associated (hand, door-knob, bed, and desk) samples. The orphaned graphs are mostly samples from one individual. As expected, common surfaces in the hand-associated realm serve as an anchor for their subgraph, connecting a number of different people, while hallway floors serve the same role for individual shoes. In contrast, orphaned graphs appear to indicate the stability of an individual's microbial signature over time and a lack of interaction with other samples.

**TABLE 2** Assortativity of metadata factors

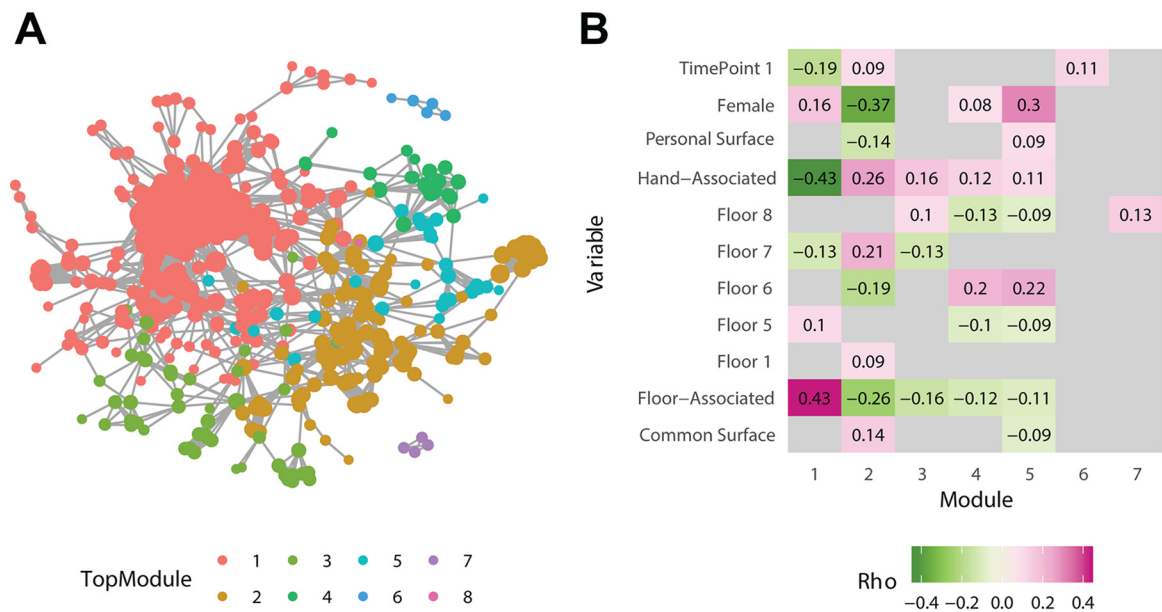
Graph	Floor	Personal vs common	Sex	Surface	Time point	Subject ID
Weighted UniFrac	0.0496	0.2062	0.0583	0.1734	0.0705	0.0687
Random forest	0.3941	0.1446	0.3932	0.1985	0.0587	0.3245

While a graph can be constructed using a beta-diversity metric (in our case weighted UniFrac distance) as above, the distance metric may not be sensitive to the microbial community of an individual. Since there is information to be gained from aggregating samples into a larger individual signature, we also constructed a graph using random forest model proximity. The proximity values from the random forest model are akin to distance and take into account the same signature used to classify individuals. It is also much sparser, as the random forest model tries to minimize distances between samples from the same individual, while keeping samples between individuals distinct. The resulting graph can be seen in Fig. S5a and S5b, where samples are colored by individual and surface type, respectively. While these graphs show rough clustering by individual, and less clustering by surface, it is unclear at what level to delineate spheres of interaction. To compare how metadata related to both the weighted UniFrac and random forest graphs, we calculated the assortativity of various metadata.

Assortativity is a metric used to quantify how often nodes in a graph attach to other similar nodes, ranging from  $-1$  to  $1$ . Positive assortativity reflects high connectivity between similar nodes, while negative assortativity indicates connections between dissimilar nodes, with zero indicating no relationship. As seen in Table 2, all metadata factors had positive assortativity, with small positive assortativity values across the weighted UniFrac graph. The highest values belonged to the identity of a given surface and its personal or common nature, which implied that similar surfaces may share similar bacteria. Time point had low assortativity across graphs, indicating that the dorm has stable signatures over time. By contrast, the random forest graph had higher assortativity measures for floor, sex, and subject identity than the weighted UniFrac graph. The random forest is trained to distinguish individuals and their signatures, and the higher subject ID assortativity showed that it was better able to connect samples from unique individuals to each other. The increased assortativity of floor and sex may be related to this, as samples from one individual were also from the same floor and the same sex. At the same time, the assortativity of floor and sex were higher than subject ID, which may indicate that their association is more than just a result of grouping by subject. In addition to metadata, we wished to understand the spheres of interaction in the dormitory.

Graph-based clustering analysis methods are often used in describing interactions in social networks. Using the Infomap clustering algorithm (46), which uses flow within a network to generate groupings, we looked at how samples clustered into spheres of interaction. The relevant scale of interactions is not always clear, and the Infomap algorithm is also hierarchical (47), allowing for clustering of samples at many scales. This allows for samples to be first classified into large clusters, known as “top modules,” and then into smaller clusters within each top module, known as submodules, which are smaller groupings of fewer samples. Using this algorithm, we identified eight top modules (Fig. 4A), with module 1 encompassing almost all shoe and floor samples. We also wished to understand how these clusters related to the samples themselves and what factors associate with this structuring. As seen in Fig. 4B, a number of factors were significantly correlated with each cluster. Cluster 1, composed mostly of shoe samples, was associated with floor samples, and negatively associated with hand samples. In contrast, modules 2 through 4 all showed association with hand samples, while module 2 was more male and more related to common spaces than other modules. Further, we wished to examine which surfaces most commonly connect these spheres of interaction, and we found that hands were significantly enriched in connections between modules compared to the larger data set (binomial test,  $P = 8.69 \times 10^{-8}$ ). Other

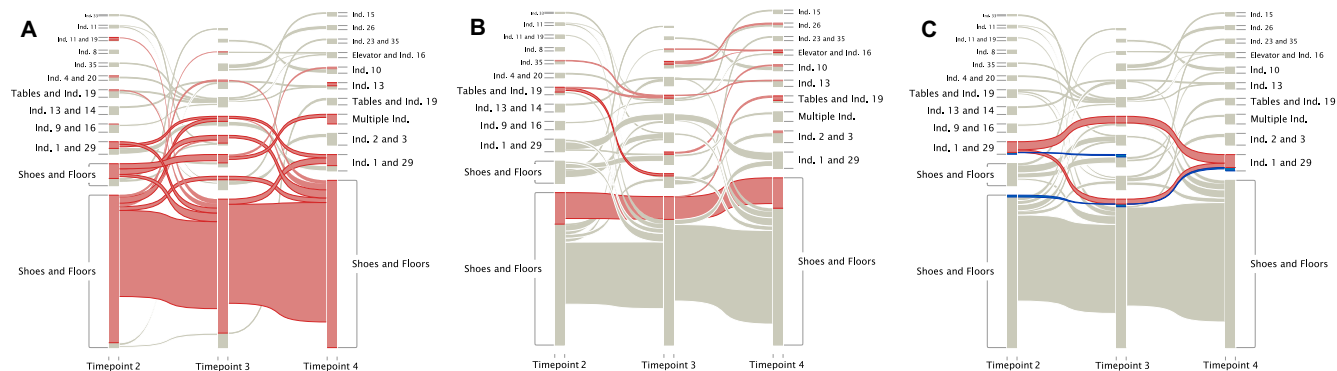




**FIG 4** (A) A graph generated using random forest model proximity scores, trained to distinguish individuals. It is thresholded by proximity less than 0.076. It is colored by Top Module, the highest-level clustering produced by Infomap. Module 1 is mostly composed of shoe and floor samples, similarly to that shown in Fig. 3. (B) Significant Spearman correlations ( $P < 0.05$ ) between each module and various metadata categories.

hand-associated surfaces showed enrichment, including common tables, doors, and bathroom doors (Fig. S6).

Of particular interest was how samples grouped over time, as samples that stably group together may indicate association. As we sampled the dorm over multiple discrete time points, we have a number of separate interaction graphs at each time point. This can be expressed as a multilayer graph (48, 49), where each time point is its own graph, representing interactions at a single point in time. These separate graphs are also connected by interactions which occurred between samplings, which we can estimate using the distances between these samples. To account for this structure, we employed a multilayer implementation of the Infomap algorithm to look at the stability of interactions over time (50). Here, we used samples from time points 2 to 4, as time point 1 consisted of only samples from participant hands and was not directly comparable to the other three time points. This is presented in Fig. 5, where samples are clustered at each time point and their membership in clusters is tracked over time. Shoe and floor samples showed high stability over time, where most samples cocluster over



**FIG 5** Alluvial diagrams depicting the clustering of samples over time. (A) All samples that were associated with the floor (hallway floors, bedroom floors, and shoes) were colored red. Ind., individual. (B) All common surfaces desks, bathroom doors, elevator buttons, and hallway floors. (C) Individual 1 (red) and individual 29 (blue) are indicated in color.



time in the same clusters (Fig. 5A). Common surfaces had a similar pattern, where common floor samples clustered consistently, while common hand-associated samples could be affiliated with different samples from many individuals (Fig. 5B). To demonstrate how two individuals freely cluster over time, we colored all the samples from two participants (individuals 1 and 29) (Fig. 5C). The samples from individual 1 show that all samples from an individual do not always cluster together, indeed, despite clustering during time points 2 and 4, they cluster separately at time point 2. In contrast, there are two sets of samples from individual 29 that cluster consistently and independently, which reflects the division of samples by type (shoe versus floor) rather than by unique individual.

## DISCUSSION

The use of human microbial signatures as trace evidence remains a young and inexact science. In order for this developing field to become a useful forensic tool, methods will need to be optimized and the myriad factors that influence our microbial interaction with built environments will need to be disentangled. Here, we compared classification methods to link residents to their rooms and personal effects in a common dormitory environment. For classifying individuals, minimum entropy decomposition (MED) was determined to be the best choice based on its high error ratio and ability to recover higher importance scores for all taxa. Further, it appears that the exact sequence variants produced by DADA2 and MED in general are better at identifying individuals than OTU-clustering methods such as UPARSE. This is unsurprising, as exact sequence variants avoid grouping closely related sequences together that could be indicative of individuals. MED is able to recover more diversity within the main skin-associated taxa from the phyla *Proteobacteria*, *Fusobacteria*, *Bacteroidetes*, and *Actinobacteria*, and those sequences are more closely related phylogenetically.

This diversity translates directly into increased utility in classification, as the sequences generated by MED have higher importance scores, and thus discriminative ability, than DADA2 or UPARSE. This is true even at the genus level, indicating that it is able to produce more individual-specific sequences within common skin-associated taxa. At the same time, MED did not recover nearly as many phyla as UPARSE or DADA2 did, and thus underestimated the full diversity of the data set. In this case, fine-scale diversity in highly abundant phyla was what we sought, but it could pose an issue for classification using highly divergent and low-abundance organisms. Further, the increased diversity that MED produces should be interpreted in light of observations that MED can produce false-positive sequence variants in data from mock communities (36, 51).

Here we show that skin-associated samples are useful in linking individuals to rooms that they have inhabited. These microbial signatures appear largely stable, as samples across all time points, spanning a period of 4 weeks, are useful in classification. Temporal stability has been observed in the skin microbiome (15, 17), and this stability in our case extends to the personal samples of each individual. Since individuals contribute their microbial signature to their environment, the presence of a roommate can interfere with classification. There is potential for interaction both between the skin of individuals, which has been seen in couples (20), and a mixing of signatures in the room itself, as seen in examples of cohabitating individuals (3, 19). Roommates were a confounding factor, and classification error linearly correlated with the number of roommates and accounted for 30% of the variance. The increase in classification error should provoke caution in those who seek to discern the signatures of cohabitating individuals, as a mixing of signatures can obscure the true inhabitant of a room.

As in most prior work with microbial forensics (9, 21), our analyses are based on 16S rRNA sequencing methods (52). In contrast, other methods rely on metagenomic markers, which are determined from the sample (22, 24) or known *a priori* (23). Metagenomic methods require shotgun metagenomic sequencing and may be difficult to implement in low-biomass samples or situations where host DNA overwhelms

bacterial DNA (53, 54). However, these methods are often significantly more accurate and represent an important direction for future research.

In addition to classification, we were able to examine the larger interaction structure of the college dormitory. The dormitory has two large spheres of interaction, structured by their association with either floors or hands. Floor samples are highly connected to each other and form a dense subgraph in both weighted UniFrac and random forest-based graphs. Close interactions among floor-associated samples may be due to the homogenizing effect of walking, and this has been observed in prior studies (9). In contrast, hand-associated samples appear to be structured both by the individual from which the sample originates and by the common surfaces they interact with. As seen in the weighted UniFrac network analysis, common surfaces form a backbone connecting many separate individuals, which identifies them as potential points of interaction. When looking at the random forest-based analysis, top modules of samples form clusters of interaction that are associated with a number of factors, including whether the samples are hand associated, which floor they originated from, and the sex of the individuals who contributed them. In addition, connections between these top modules are enriched for hands, indicating that hands may be points of interactions between individuals.

When looking at interaction networks, we found that sample clustering over time was highly dependent on sample type, with floor-associated samples showing long, persistent interactions, while common surfaces were more freely interacting. Individual signatures do not always cluster together over time and can form associations over time. While other sources of data, including sexual partners (55) and coauthorship (56), have been used to analyze networks of human interaction (57), this is the first study of which the authors are aware to identify networks of interaction using microbial signatures.

Through individual and common space sampling of a college dormitory, we discovered that MED and other sequence-based methods are superior to those that rely upon OTU clustering. We have also found that common surfaces form a scaffold connecting many individuals, and further, that spheres of interaction are disproportionately connected by hands, indicating them as a means of transmission of microbes. Finally, we have characterized the persistence of interaction and found differences in persistence based on the sample type.

## MATERIALS AND METHODS

**Study design and sample collection.** We collected personal samples from 37 participants in 28 distinct dorm rooms (see Table S1 and S2 in the supplemental material). Samples were collected by swabbing a sterile cotton BD-Swube applicator against the dry surface of interest. Sampling kits were given to study participants for self-sampling with instructions. For the first time point, only the hands of individuals were sampled. The desk, floor, fitted bed sheet, and interior doorknobs of each participant's room, along with the dominant hand and shoe of the participant, were sampled at three time points after the first time point. The first time point occurred before occupants left for a scheduled school break (end of a quarter) and then immediately upon return. The third and fourth time points were taken 2 and 4 weeks after spring break.

Participants also completed a questionnaire which collected basic information on the subject, the conditions specific to their dorm room, and who they interacted with in their dorm room during the sampling period. This questionnaire was completed each time a set of samples was collected.

Common surfaces were also sampled similarly. Common surfaces specific to the 5th floor included tables in the dormitory lounge, and the handle of the entry door to the lounge. On each floor of the dormitory, the door handles of bathrooms, the floors of each hallway, and the elevator buttons were sampled. Each floor had its own unique combination, and these were swabbed at the same time as personal surfaces.

**Sample processing.** DNA was extracted from each sample using a low biomass variation of the MO BIO Powersoil DNA extraction protocol. 16S rRNA was amplified with the Earth Microbiome Project 16S Illumina Amplicon Protocol (<http://www.earthmicrobiome.org/protocols-and-standards/16s/>). The V4 region of the 16S rRNA gene was targeted with the 515F-806RB primer pair. Sequencing was performed using an Illumina Miseq sequencer with the protocol described by Caporaso et al. (52).

**Sequence processing.** Each method was processed using the default workflows provided in reference papers given below.

(i) **UPARSE.** Demultiplexed sequences were merged using vsearch v2.3.0 (58) with 10,040,708 successful paired-end reads merged together. Sequences were quality filtered with a maximum expected error of 0.5, with 9,057,613 remaining sequences. Sequences were then dereplicated for 1,276,202 unique sequences. Sequences were then clustered at 97% identity, with 11,658 OTUs and 42,539 chimeras. Sequences were then matched to OTUs with 93.28% of sequences matched to OTUs. A total of 6,011

OTUs passed sequence processing. Chloroplast and mitochondrial DNA was removed, and samples were rarefied to 4,000 counts per sample.

**(ii) MED.** Sequences were processed according to the methods described by Eren et al. (35). Demultiplexed paired-end reads were merged using *illumina-utils* (59), with Q30 check imposed on sequences, leading to 10,023,266 successfully merged out of 10,023,266 reads. Gaps between sequences were padded with blanks, and samples were decomposed using a -M of 100. A total of 1,732,615 outliers were removed by quality control, and remaining sequences were sorted into 3,748 nodes after refinement. A total of 3,352 passed quality control. Chloroplast and mitochondrial DNA was removed, and samples were rarefied to 4,000 counts per sample.

**(iii) DADA2.** The filtering step of DADA2 version 1.03 was run with no ambiguous base (maxN of 0), maximum expected errors of 2, and quality of truncation of 2. All other commands were run on default settings. Sequences were merged after performing quality filtering. After merging, 34,043 sequences were observed, and 18,329 sequences were not chimeras. A total of 4,307 unique sequences passed final quality filtering. Chloroplast and mitochondrial DNA was removed, and samples were rarefied to 4,000 counts per sample.

**Taxonomic identification.** All sequences were taxonomically identified using the same implementation of RDP (60) implemented in DADA2 to enable comparison between the sequencing methods. Taxonomy was assigned using the SILVA (61) training set version 123.

**Phylogenetic trees.** Sequences were aligned with the R package *MSA* (62) version 1.4.5, using the Muscle (63, 64) algorithm. Phylogenetic trees were then generated using the R package *Phangorn* (65) version 2.1.1. The tree was first created by neighbor joining and fitted with GTR clock model.

**Data analysis and visualization.** Data cleaning and shaping were performed using R 3.3.2-R3.5.2 and the packages *dplyr* 0.7.8 (66) and *reshape2* 1.4.3 (67). Visualization and analysis were performed using *phyloseq* 1.26 (68), *igraph* v1.2.4.1 (69), *ggnetwork* (70), and *ggplot2* v3.1.1 (71). Boxplots with significance were generated using *ggpubr* v0.2 (72). Phylogenetic distance was calculated using the “cophenetic.phylo” function from *ape* v5.3 (73). Differential abundance calculations were performed using DESeq2 v1.12.4 (74). Diversity measures were calculated using *vegan* v2.5-4 (75). Ideas for analysis, along with basic code snippets were taken from Callahan et al. (76). Community clustering was performed using the Infomap (77, 78) and alluvial diagrams were generated using the “Map & Alluvial Generator” (<http://www.mapequation.org/apps/MapGenerator.html>).

**Random forests.** Random forest models were generated using *randomForest* 4.6-13 (79) and *ranger* v0.10 (80). For classification of individuals, all room samples for individuals were used to predict the hands of individuals. This was repeated 10 times for cross validation (CV-10), and proximities/importance scores were averaged across runs. For comparisons of men versus women, all personal samples were subdivided into testing/training sets, with two thirds of samples in the training set and one third in the testing set. This was run thrice (CV-3).

**Ethics.** This study (institutional review board [IRB] number IRB15-0373) was approved by Biological Sciences Division (BSD) IRB Committee A of The University of Chicago Biological Sciences Division/ University of Chicago Medical Center.

**Data availability.** Sequencing data and sample data are available from Qiita, study ID 12470, and from EBI, project number [PRJEB33050/ERP115809](https://www.ebi.ac.uk/ena/browser/study/PRJEB33050/ERP115809). Phyloseq object files from each of the three sequencing methods, along with sequence tables, taxonomy tables, anonymized sample data, and phylogenetic trees are available on Github at <https://github.com/MiPZR/Dorm-Microbiome>.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mBio.01054-19>.

**FIG S1**, PDF file, 0.01 MB.

**FIG S2**, PDF file, 0.01 MB.

**FIG S3**, PDF file, 0.8 MB.

**FIG S4**, TIF file, 0.1 MB.

**FIG S5**, PDF file, 0.2 MB.

**FIG S6**, PDF file, 0.01 MB.

**TABLE S1**, DOCX file, 0.01 MB.

**TABLE S2**, DOCX file, 0.01 MB.

## ACKNOWLEDGMENTS

This work was sponsored by National Institutes of Justice award 2015-DN-BX-K430. We thank Sophia Weaver for indispensable assistance in recruiting individuals.

## REFERENCES

- Dunn RR, Fierer N, Henley JB, Leff JW, Menninger HL. 2013. Home life: factors structuring the bacterial diversity found within and between homes. *PLoS One* 8:e64133. <https://doi.org/10.1371/journal.pone.0064133>.
- Flores GE, Bates ST, Caporaso JG, Lauber CL, Leff JW, Knight R, Fierer N. 2013. Diversity, distribution and sources of bacteria in residential kitchens. *Environ Microbiol* 15:588–596. <https://doi.org/10.1111/1462-2920.12036>.
- Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S, Metcalf JL, Ursell LK,

- Vázquez-Baeza Y, Treuren WV, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, Gilbert JA. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* 345: 1048–1052. <https://doi.org/10.1126/science.1254529>.
4. Miletto M, Lindow SE. 2015. Relative and contextual contribution of different sources to the composition and abundance of indoor air bacteria in residences. *Microbiome* 3:61. <https://doi.org/10.1186/s40168-015-0128-z>.
  5. Meadow JF, Altrichter AE, Kembel SW, Kline J, Mhuireach G, Moriyama M, Northcutt D, O'Connor TK, Womack AM, Brown GZ, Green JL, Bohannan BJM. 2014. Indoor airborne bacterial communities are influenced by ventilation, occupancy, and outdoor air source. *Indoor Air* 24:41–48. <https://doi.org/10.1111/ina.12047>.
  6. Poza M, Gayoso C, Gómez MJ, Rumbo-Feal S, Tomás M, Aranda J, Fernández A, Bou G. 2012. Exploring bacterial diversity in hospital environments by GS-FLX titanium pyrosequencing. *PLoS One* 7:e44105. <https://doi.org/10.1371/journal.pone.0044105>.
  7. Kembel SW, Jones E, Kline J, Northcutt D, Stenson J, Womack AM, Bohannan BJ, Brown GZ, Green JL. 2012. Architectural design influences the diversity and structure of the built environment microbiome. *ISME J* 6:1469–1479. <https://doi.org/10.1038/ismej.2011.211>.
  8. Wood M, Gibbons SM, Lax S, Shoo-Anton TW, Owens SM, Kennedy S, Gilbert JA, Hampton-Marcell JT. 2015. Athletic equipment microbiota are shaped by interactions with human skin. *Microbiome* 3:25. <https://doi.org/10.1186/s40168-015-0088-3>.
  9. Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen JA, Gilbert JA. 2015. Forensic analysis of the microbiome of phones and shoes. *Microbiome* 3:21. <https://doi.org/10.1186/s40168-015-0082-9>.
  10. Hoisington A, Maestre JP, Kinney KA, Siegel JA. 2016. Characterizing the bacterial communities in retail stores in the United States. *Indoor Air* 26:857–868. <https://doi.org/10.1111/ina.12273>.
  11. Lax S, Sangwan N, Smith D, Larsen P, Handley KM, Richardson M, Guyton K, Krezalek M, Shogan BD, Defazio J, Flemming I, Shakhsher B, Weber S, Landon E, Garcia-Houchins S, Siegel J, Alverdy J, Knight R, Stephens B, Gilbert JA. 2017. Bacterial colonization and succession in a newly opened hospital. *Sci Transl Med* 9:eaah6500. <https://doi.org/10.1126/scitranslmed.aah6500>.
  12. Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown GZ, Green JL, Bohannan B. 2015. Humans differ in their personal microbial cloud. *PeerJ* 3:e1258. <https://doi.org/10.7717/peerj.1258>.
  13. Qian J, Hospodsky D, Yamamoto N, Nazaroff WW, Peccia J. 2012. Size-resolved emission rates of airborne bacteria and fungi in an occupied classroom. *Indoor Air* 22:339–351. <https://doi.org/10.1111/j.1600-0668.2012.00769.x>.
  14. Adams RI, Bhargava S, Pasut W, Arens EA, Taylor JW, Lindow SE, Nazaroff WW, Bruns TD. 2015. Chamber bioaerosol study: outdoor air and human occupants as sources of indoor airborne microbes. *PLoS One* 10: e0128022. <https://doi.org/10.1371/journal.pone.0128022>.
  15. Oh J, Byrd AL, Park M, Kong HH, Segre JA. 2016. Temporal stability of the human skin microbiome. *Cell* 165:854–866. <https://doi.org/10.1016/j.cell.2016.04.008>.
  16. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. *Science* 326:1694–1697. <https://doi.org/10.1126/science.1177486>.
  17. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, NISC Comparative Sequencing Program, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009. Topographical and temporal diversity of the human skin microbiome. *Science* 324:1190–1192. <https://doi.org/10.1126/science.1171700>.
  18. Hospodsky D, Qian J, Nazaroff WW, Yamamoto N, Bibby K, Rismani-Yazdi H, Peccia J. 2012. Human occupancy as a source of indoor airborne bacteria. *PLoS One* 7:e34867. <https://doi.org/10.1371/journal.pone.0034867>.
  19. Song SJ, Lauber C, Costello EK, Lozupone CA, Humphrey G, Berg-Lyons D, Caporaso JG, Knights D, Clemente JC, Nakielnny S, Gordon JI, Fierer N, Knight R. 2013. Cohabiting family members share microbiota with one another and with their dogs. *Elife* 2:e00458. <https://doi.org/10.7554/eLife.00458>.
  20. Ross AA, Doxey AC, Neufeld JD. 2017. The skin microbiome of cohabiting couples. *mSystems* 2:e00043-17. <https://doi.org/10.1128/mSystems.00043-17>.
  21. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A* 107:6477–6481. <https://doi.org/10.1073/pnas.1000162107>.
  22. Schmedes SE, Woerner AE, Budowle B. 2017. Forensic human identification using skin microbiomes. *Appl Environ Microbiol* 83:e01672-17. <https://doi.org/10.1128/AEM.01672-17>.
  23. Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. 2018. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. *Forensic Sci Int Genet* 32:50–61. <https://doi.org/10.1016/j.fsigen.2017.10.004>.
  24. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci U S A* 112:E2930–E2938. <https://doi.org/10.1073/pnas.1423854112>.
  25. Maestre JP, Jennings W, Wylie D, Horner SD, Siegel J, Kinney KA. 2018. Filter forensics: microbiota recovery from residential HVAC filters. *Microbiome* 6:22. <https://doi.org/10.1186/s40168-018-0407-6>.
  26. Hewitt KM, Gerba CP, Maxwell SL, Kelley ST. 2012. Office space bacterial abundance and diversity in three metropolitan areas. *PLoS One* 7:e37849. <https://doi.org/10.1371/journal.pone.0037849>.
  27. Kembel SW, Meadow JF, O'Connor TK, Mhuireach G, Northcutt D, Kline J, Moriyama M, Brown GZ, Bohannan BJM, Green JL. 2014. Architectural design drives the biogeography of indoor bacterial communities. *PLoS One* 9:e87093. <https://doi.org/10.1371/journal.pone.0087093>.
  28. Huse SM, Welch DM, Morrison HG, Sogin ML. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 12:1889–1898. <https://doi.org/10.1111/j.1462-2920.2010.02193.x>.
  29. Patin NV, Kunin V, Lidström U, Ashby MN. 2013. Effects of OTU clustering and PCR artifacts on microbial diversity estimates. *Microb Ecol* 65: 709–719. <https://doi.org/10.1007/s00248-012-0145-4>.
  30. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102: 2567–2572. <https://doi.org/10.1073/pnas.0409727102>.
  31. Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10:996–998. <https://doi.org/10.1038/nmeth.2604>.
  32. Nguyen N-P, Warnow T, Pop M, White B. 2016. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* 2:16004. <https://doi.org/10.1038/npjbiofilms.2016.4>.
  33. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11:2639–2643. <https://doi.org/10.1038/ismej.2017.119>.
  34. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* 4:1111–1119. <https://doi.org/10.1111/2041-210X.12114>.
  35. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* 9:968–979. <https://doi.org/10.1038/ismej.2014.195>.
  36. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>.
  37. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. <https://doi.org/10.1128/mSystems.00191-16>.
  38. Eren AM, Borisy GG, Huse SM, Welch J. 2014. Oligotyping analysis of the human oral microbiome. *Proc Natl Acad Sci U S A* 111:E2875–E2884. <https://doi.org/10.1073/pnas.1409644111>.
  39. Fisher JC, Eren AM, Green HC, Shanks OC, Morrison HG, Vineis JH, Sogin ML, McLellan SL. 2015. Comparison of sewage and animal fecal microbiomes by using oligotyping reveals potential human fecal indicators in multiple taxonomic groups. *Appl Environ Microbiol* 81:7023–7033. <https://doi.org/10.1128/AEM.01524-15>.
  40. Schmidt VT, Reveillaud J, Zettler E, Mincer TJ, Murphy L, Amaral-Zettler LA. 2014. Oligotyping reveals community level habitat selection within the genus *Vibrio*. *Front Microbiol* 5:563. <https://doi.org/10.3389/fmicb.2014.00563>.
  41. Berry MA, White JD, Davis TW, Jain S, Johengen TH, Dick GJ, Sarnelle O, Denef VJ. 2017. Are oligotypes meaningful ecological and phylogenetic units? A case study of *Microcystis* in freshwater lakes. *Front Microbiol* 8:365. <https://doi.org/10.3389/fmicb.2017.00365>.
  42. Eren AM, Sogin ML, Morrison HG, Vineis JH, Fisher JC, Newton RJ,



- McLellan SL. 2015. A single genus in the gut microbiome reflects host preference and specificity. *ISME J* 9:90–100. <https://doi.org/10.1038/ismej.2014.97>.
43. Shi T, Horvath S. 2006. Unsupervised learning with random forest predictors. *J Comput Graph Stat* 15:118–138. <https://doi.org/10.1198/106186006X94072>.
  44. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. *FEMS Microbiol Rev* 35:343–359. <https://doi.org/10.1111/j.1574-6976.2010.00251.x>.
  45. Luongo JC, Barberán A, Hacker-Cary R, Morgan EE, Miller SL, Fierer N. 2017. Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. *Indoor Air* 27:338–344. <https://doi.org/10.1111/ina.12302>.
  46. Rosvall M, Bergstrom CT. 2008. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A* 105:1118–1123. <https://doi.org/10.1073/pnas.0706851105>.
  47. Rosvall M, Bergstrom CT. 2011. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One* 6:e18209. <https://doi.org/10.1371/journal.pone.0018209>.
  48. De Domenico M, Solé-Ribalta A, Cozzo E, Kivelä M, Moreno Y, Porter MA, Gómez S, Arenas A. 2013. Mathematical formulation of multilayer networks. *Phys Rev X* 3:041022.
  49. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. 2014. Multilayer networks. *J Complex Netw* 2:203–271. <https://doi.org/10.1093/comnet/cnu016>.
  50. Aslak U, Rosvall M, Lehmann S. 2018. Constrained information flows in temporal networks reveal intermittent communities. *Phys Rev E* 97:062312. <https://doi.org/10.1103/PhysRevE.97.062312>.
  51. Caruso V, Song X, Asquith M, Karstens L. 2019. Performance of microbiome sequence inference methods in environments with varying biomass. *mSystems* 4:e00163-18. <https://doi.org/10.1128/mSystems.00163-18>.
  52. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6:1621–1624. <https://doi.org/10.1038/ismej.2012.8>.
  53. Sharpston TJ. 2014. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci* 5:209. <https://doi.org/10.3389/fpls.2014.00209>.
  54. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 35:833–844. <https://doi.org/10.1038/nbt.3935>.
  55. Liljeros F, Edling CR, Amaral LAN, Stanley HE, Åberg Y. 2001. The web of human sexual contacts. *Nature* 411:907–908. <https://doi.org/10.1038/35082140>.
  56. Barabási AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T. 2002. Evolution of the social network of scientific collaborations. *Phys Stat Mech Its Appl* 311:590–614. [https://doi.org/10.1016/S0378-4371\(02\)00736-7](https://doi.org/10.1016/S0378-4371(02)00736-7).
  57. Palla G, Barabási A-L, Vicsek T. 2007. Quantifying social group evolution. *Nature* 446:664–667. <https://doi.org/10.1038/nature05670>.
  58. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. <https://doi.org/10.7717/peerj.2584>.
  59. Eren AM, Vineis JH, Morrison HG, Sogin ML. 2013. A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS One* 8:e66643. <https://doi.org/10.1371/journal.pone.0066643>.
  60. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267. <https://doi.org/10.1128/AEM.00062-07>.
  61. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Priesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res* 42:D643–D648. <https://doi.org/10.1093/nar/gkt1209>.
  62. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. 2015. msa: an R package for multiple sequence alignment. *Bioinformatics* 31:3997–3999. <https://doi.org/10.1093/bioinformatics/btv494>.
  63. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. <https://doi.org/10.1093/nar/gkh340>.
  64. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. <https://doi.org/10.1186/1471-2105-5-113>.
  65. Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593. <https://doi.org/10.1093/bioinformatics/btq706>.
  66. Wickham H, François R, Henry L, Müller K. 2018. dplyr: a grammar of data manipulation.
  67. Wickham H. 2007. Reshaping data with the reshape package. *J Stat Softw* 21:1–20. <https://doi.org/10.18637/jss.v021.i12>.
  68. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. <https://doi.org/10.1371/journal.pone.0061217>.
  69. Csardi G, Nepusz T. 2005. The igraph software package for complex network research. *InterJournal Complex Syst* 1695:9.
  70. Briatte F. 2016. ggnet: geometries to plot networks with “ggplot2.”
  71. Wickham H. 2011. ggplot2. *WIREs Comp Stat* 3:180–185. <https://doi.org/10.1002/wics.147>.
  72. Kassambara A. 2018. ggpubr: “ggplot2” based publication ready plots.
  73. Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290. <https://doi.org/10.1093/bioinformatics/btg412>.
  74. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550. <https://doi.org/10.1186/s13059-014-0550-8>.
  75. Dixon P, Palmer MW. 2003. VEGAN, a package of R functions for community ecology. *J Veg Sci* 14:927–930. <https://doi.org/10.1111/j.1654-1103.2003.tb02228.x>.
  76. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. 2016. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Res* 5:1492. <https://doi.org/10.12688/f1000research.8986.2>.
  77. Rosvall M, Axelsson D, Bergstrom CT. 2009. The map equation. *Eur Phys J Spec Top* 178:13–23. <https://doi.org/10.1140/epjst/e2010-01179-1>.
  78. Bohlin L, Edler D, Lancichinetti A, Rosvall M. 2014. Community detection and visualization of networks with the map equation framework, p 3–34. In Ding Y, Rousseau R, Wolfram D (ed), *Measuring scholarly impact: methods and practice*. Springer International Publishing, Cham, Switzerland.
  79. Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* 2/3:18–22.
  80. Wright MN, Ziegler A. 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Software* 77:1–17. <https://doi.org/10.18637/jss.v077.i01>.